四膜虫基因表达数据库: 四膜虫全基因组基因表达芯片和协同表达分析的数据资源

熊杰^{①②†}, 陆星亦^{①②†}, 陆钰明^{①†}, 曾宏辉^{①†}, 袁冬霞^①, 冯立芳^①, 畅悦^{①②},
BOWEN Josephine^③, GOROVSKY Martin^③, 傅诚杰^④, 缪炜^{①④*}

① 中国科学院水生生物研究所水生生物多样性与保护重点实验室, 武汉 430072;

② 中国科学院研究生院, 北京 100049;

③ Department of Biology, University of Rochester, Rochester, NY 14627, USA;

④ 中国科学院水生生物研究所, 淡水生态与生物技术国家重点实验室, 武汉 430072

† 同等贡献

* 联系人, E-mail: miaowei@ihb.ac.cn

收稿日期: 2010-07-05; 接受日期: 2010-10-12

国家自然科学基金(批准号: 30870356 和 30970424)、中国科学院知识创新工程(批准号: KSCX2-YW-G-072)和 U.S. National Institutes of Health(批准号: GM021793 和 GM072752)资助项目

摘要 嗜热四膜虫是一种优良的真核模式生物, 对其功能基因组学的分析将为研究细胞和分子生物学的基础性重大问题提供新的机遇。四膜虫基因表达数据库(TGED)是原生动物纤毛虫中第一个基因表达数据库, 整个数据库包括了四膜虫 3 个重要的生理和发育阶段 20 个时间点的全基因组基因表达数据, 提供了友好的网页界面(<http://tged.ihb.ac.cn>)供用户获取这些数据。通过基因的标识号或描述信息, 用户可以获取基因表达谱信息和协同表达基因。此外, TGED 同时提供了制备四膜虫基因表达芯片的样品制备方法。欢迎研究者上传和分享其所有的基因芯片数据, 以期四膜虫或纤毛虫研究提供重要的功能基因组学资源。

关键词

嗜热四膜虫
基因表达数据库
协同表达

纤毛类原生动物四膜虫(*Tetrahymena*)自 1923 年被诺贝尔奖获得者 Andre Lwoff 纯培养以来, 已成为一种优良的真核模式生物, 被用于包括分别获得 1989 和 2009 年诺贝尔奖的核酶^[1]与端粒和端粒酶^[2]的发现。在不同的四膜虫种中, 被研究得最多的是嗜热四膜虫(*T. thermophila*), 其在标准的微生物培养技术下很容易进行培养, 不需要任何特殊的装置^[3], 并且很容易在其生活史的任意阶段进行遗传操作, 包括基因敲除和基因过表达^[4]。

嗜热四膜虫在分类上属于具泡枝的纤毛门, 与

人们所熟知的草履虫同属一个门。与其他很多纤毛虫类似, 嗜热四膜虫同样呈现出核的多态性。嗜热四膜虫的细胞通常有两个核: 一个小核和一个大核^[3]。小核包含 5 条染色体。在嗜热四膜虫接合生殖期, 来自不同细胞的减数分裂后的小核相互交换后与另一细胞中的小核融合成合子, 而嗜热四膜虫的大核则产生于合子进行有丝分裂后产生的多个细胞核中的一个^[5]。四膜虫的大核来源于小核, 大约含有 225 条染色体, 具有转录活性, 通常被称为营养核^[6]。2006 年, 嗜热四膜虫大核基因组测序完成^[6], 并相继建立

英文版见: Xiong J, Lu X Y, Lu Y M, et al. *Tetrahymena* Gene expression database (TGED): a resource of microarray data and co-expression analyses for *Tetrahymena*. *Sci China Life Sci*, 2011, 54: 65—67, doi: 10.1007/s11427-010-4114-1

了嗜热四膜虫大核基因组数据库(tetrahymena genome database, TGD)^[7]。随着分子遗传学技术和基因芯片平台的发展和建立^[8], 对嗜热四膜虫功能基因组的分析将为细胞和分子生物学基础性重大问题的研究提供新的机遇。

四膜虫基因表达数据库(TGED)的建立就是以四膜虫生长、饥饿和接合生殖 3 个时期基因芯片表达数据为基础, 提供对嗜热四膜虫基因的表达谱和协同表达基因的信息。此外, TGED 还提供了制备嗜热四膜虫基因芯片的样品、实验方法及国际同行共享基因芯片数据的平台。

1 TGED 主要内容

TGED 整合了嗜热四膜虫 50 个全基因组基因芯片表达数据^[8], 包括嗜热四膜虫 3 个重要生理和发育时期的 20 个时间点: 3 个生长时期、7 个饥饿时期和 10 个接合生殖时期。

嗜热四膜虫全基因组基因芯片平台的设计是以 TGD 中 2006 版基因注释为基础采用罗氏 NimbleGen 系统完成的。阵列结果用罗氏 NimbleScan 软件进行扫描, 对每一个可读框探针表达值的均一化采用了 RMA (robust multiarray average)方法^[9]。经过此方法处理后的值即为表达值(RMA calls), 这些表达值数据被用来进行下一步分析。将已提交 GEO (gene expression omnibus)的所有四膜虫生长、饥饿和接合生殖时期的 RMA calls 数据输入 ArrayStar (DNASTAR, Inc, Madison, WI)软件进行整合, 然后输出整合的数据文件。这个文件则为 TGED 数据库, 共包含嗜热四膜虫 3 个生理和发育时期 20 个时间点的表达数据。每个基因 20 个时间点组成的表达谱在 TGED 中通过图形显示。

由于在相同生物过程中行使功能的基因很可能表现出一致的表达谱, 基因协同表达的数据可以为分析基因间调节关系和蛋白与蛋白相互作用等提供非常有价值的信息^[10]。基于此, 使用嗜热四膜虫生长、饥饿和接合生殖的表达数据, 利用 C++编译的程序(图 1)对每个基因对的皮尔逊相关系数(Pearson correlation coefficients, R)进行计算。 $R \geq 0.8$ 的基因对被认为是潜在的协同表达基因。利用这些基因协同表达的数据来寻找四膜虫中处于蛋白复合体或同一生物过程(如四膜虫接合生殖时期的 DNA elimination

过程)中的基因被证明是非常有效的^[8]。

2 TGED 网页界面

TGED 可通过友好的网页界面(<http://tged.ihb.ac.cn/>)登录, 而且可以通过以下两种方式对数据库进行搜索: 基因 ID 和基因描述信息。

对于使用基因 ID, TGED 中的“SEARCH”页面允许用户输入一个嗜热四膜虫 2006 年预测版本(ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_thermophila/annotation_dbs/interim_annotation_release_08313006/)的基因 ID 进行搜索, 并获得基因的详细信息、表达谱信息和协同表达的基因目录(图 1)。基因的详细信息包括基因 ID(超链接到 NCBI)、序列 ID(基因芯片设计时产生的具有唯一性的 ID)、基因类型(如蛋白编码、非编码或转运 RNA)、基因产物名称(基因描述信息)和序列信息(包括 cDNA 和氨基酸序列以及基因芯片设计产生的 13~14 个探针的序列)(图 1(B)(a))。表达谱信息则通过图形显示出嗜热四膜虫 3 个生理和发育时期 20 个时间点的表达数据(图 1(B)(b))。而协同表达基因则包含了与搜索基因具有相似表达谱的、且相关系数大于 0.8 的候选基因(图 1(B)(c)), 用户可以通过点击查看任意候选基因的表达谱及其协同表达基因信息。

TGED 也可以通过关键词进行搜索, 用户输入的关键词会被用来与嗜热四膜虫 2006 版预测基因注释信息进行比较, TGED 会列出所有包含此关键词的基因。通过选择用户关注的基因, 用户同样可以获取基因的详细信息、表达谱信息和协同表达基因信息。

3 TGED 其他服务

TGED 中的“Sample Preparation”页面提供了四膜虫生长、饥饿和接合生殖下的标准培养方法, 使用该方法对基因芯片实验所需的 RNA 进行制备后通过 TGED 提供的标准 Bioanalyzer 分析谱对 RNA 质量进行检测。“Submit”和“Download”页面为国际同行提供交流和分享其四膜虫基因芯片数据的平台。用户可以在“Submit”界面提交其所拥有的四膜虫基因芯片的“title”, “summary”, “overall design”和“contributor”信息, 也可以在“Download”页面通过从管理员处所获得 ID 和密码对数据进行下载。

图 1 以 THERM_00000010 为例进行搜索的 TGED 网页界面
(A) 页面顶部; (B) 搜索结果. (a) 基因详细信息; (b) 基因表达谱信息; (c) 协同表达基因信息

4 未来发展

未来, TGED 将收集、整合和上载更多的四膜虫基因表达数据, 如环境污染暴露下的全基因组基

因芯片表达数据和特定基因敲除后的表达数据. 另外, 还将利用更多的算法进行协同表达基因分析, 进而构建四膜虫的代谢通路、蛋白作用和基因调控网络.

参考文献

- Kruger K, Grabowski P J, Zaug A J, et al. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. Cell, 1982, 31: 147—157
- Greider C W, Blackburn E H. Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. Cell, 1985, 43: 405—413
- Asai D J, James F J, Leslie W, et al. *Tetrahymena thermophila*, vol 62. Salt Lake City: Academic Press, 2000
- Turkewitz A P, Orias E, Kapler G. Functional genomics: the coming of age for *Tetrahymena thermophila*. Trends Genet, 2002, 18: 35—40
- Martindale D W, Allis C D, Bruns P J. Conjugation in *Tetrahymena thermophila*. A temporal analysis of cytological stages. Exp Cell Res, 1982, 140: 227—236
- Eisen J A, Coyne R S, Wu M, et al. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol, 2006, 4: e286
- Stover N A, Krieger C J, Binkley G, et al. Tetrahymena Genome Database(TGD): a new genomic resource for *Tetrahymena thermophila* research. Nucleic Acids Res, 2006, 34(Database issue): D500—503
- Miao W, Xiong J, Bowen J, et al. Microarray analyses of gene expression during the *Tetrahymena thermophila* life cycle. PLoS One, 2009, 4: e4429
- Irizarry R A, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 2003, 4: 249—264
- Bhardwaj N, Lu H. Correlation between gene expression profiles and protein-protein interactions within and across genomes. Bioinformatics, 2005, 21: 2730—2738